CHAPTER 5

Statistical Mechanics

5.1. Mechanics

We begin the discussion of statistical mechanics by a quick review of standard mechanics.

Suppose we are given N particles whose position coordinates are given by a set of scalar quantities $q_1, \dots, q_n$. In a d dimensional space one needs d numbers to specify a location, so that $n = Nd$. The rate of change of the position is

$$\frac{d}{dt} q_i = \dot{q}_i .$$

(This dot notation for the time derivative goes back to Newton and makes some of the formulas below look less cluttered.) A good way to write down the laws of motion is to specify a Lagrangian $\mathcal{L} = \mathcal{L}(q_i, \dot{q}_i, t)$ and follow the steps that will now be described; this procedure can be used for laws other than those of Newtonian mechanics as well. For any path $q(s)$, $t_0 \le s \le t$, that could take the particles from their locations at time t_0 to their locations at time t, we define an "action" by

$$A = \int_{t_0}^{t} \mathcal{L}(q(s), \dot{q}(s), s) ds,$$

and we require that the motion (according to the mechanics embodied in the Lagrangian) that takes us from $q(t_0)$ to $q(t)$ be along a path which is an extremal of the action. In other words, for the motion described by the functions $q(t)$ to obey the physics in the Lagrangian, it has to be such that perturbing it a little, say from $q(t)$ to $q(t) + \delta q(t)$, changes the action $A = \int_{t_0}^{t} \mathcal{L} ds$ very little. We simplify the analysis here by assuming that $\mathcal{L}$ does not explicitly depend on t. Then

$$\begin{aligned} \delta A = \delta \int_{t_0}^{t} \mathcal{L}(q, \dot{q}) ds &= \int_{t_0}^{t} \left(\mathcal{L}(q + \delta q, \dot{q} + \delta \dot{q}) - \mathcal{L}(q, \dot{q}) \right) ds \\ &= 0 + O(\delta q^2, \delta \dot{q}^2), \end{aligned}$$

where

$$\mathcal{L}(q+\delta q, \dot{q}+\delta\dot{q}) = \mathcal{L}(q_i, \dot{q}_i) + \sum \delta q_i \frac{\partial \mathcal{L}}{\partial q_i} + \sum \delta \dot{q}_i \frac{\partial \mathcal{L}}{\partial \dot{q}_i} + 0(\delta q^2, \delta \dot{q}^2).$$

By integration by parts we find

$$\begin{aligned}\delta \int_{t_0}^{t} \mathcal{L}\, ds &= \int_{t_0}^{t} \left(\sum \delta q_i \frac{\partial \mathcal{L}}{\partial q_i} + \sum \delta \dot{q}_i \frac{\partial \mathcal{L}}{\partial \dot{q}_i} + O(\delta q^2, \delta \dot{q}^2) \right) ds \\ &= \int_{t_0}^{t} \left(\sum \delta q_i \left(\frac{\partial \mathcal{L}}{\partial q_i} - \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_i} ds \right) + O(\delta q^2) \right) ds.\end{aligned}$$

For the path $q(t)$ to be extremal the first term has to vanish, and we conclude that

$$\frac{\partial \mathcal{L}}{\partial q_i} - \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_i} = 0,$$

for all $i = 1, \ldots, n$. These are the Lagrange equations of motion.

Example. Change notation so that $x = q$, $\dot{x} = \dot{q}$, and think of x as a coordinate in a one dimensional space. Assume that a particle of mass m at x is acted on by a force F of the form $F = -\nabla V$, where $V = V(x)$ is a potential. Specify the laws of motion by setting $\mathcal{L} = \frac{1}{2} m \dot{x}^2 - V(x)$. The Lagrange equation of motion is

$$\frac{\partial \mathcal{L}}{\partial x} - \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}} = 0$$

or equivalently

$$-\frac{\partial V}{\partial x} - \frac{d}{dt}(m\dot{x}) = 0,$$

which is Newton's second law, $F = m\ddot{x}$.

This formalism is also useful in quantum mechanics, where the probability density of going from $q(t_0)$ to $q(t)$ is the square of the path integral

$$\int e^{-2\pi i A/h}\, \mathcal{D}(q).$$

Here h is Planck's constant and $\int \mathcal{D}(q)$ stands for integration over all paths that connect $q(t_0)$ to $q(t)$. As $h \to 0$ one gets back the variational principle of the beginning of this section. Note that in the physicists' notation for path integrals (Chapter 3), the Lagrangian appeared in the argument of the exponential, so that modulo the insertion of the factor $2\pi i/h$ the quantum mechanical path integral becomes a path integral roughly in the sense of Chapter 3.

We shall use the equations of motion mostly in their Hamiltonian form: Define a momentum p_i conjugate to q_i by $p_i = \partial\mathcal{L}/\partial\dot{q}_i$. The Hamiltonian function is

$$H = \sum p_i \dot{q}_i - \mathcal{L},$$

and the equations of motion can be written as

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q_i}. \tag{5.1}$$

The proof that these equations are equivalent to the Lagrangian equations is just a manipulation of differentials which we leave to the reader.

Example. Let $\mathcal{L} = \frac{1}{2}m\dot{x}^2 - V(x)$ as before, with $q = x$. Then $p = m\dot{x}$ and

$$H = p\dot{q} - \mathcal{L} = (m\dot{x})\dot{x} - \left(\frac{1}{2}m\dot{x^2} - V(x)\right) = \frac{1}{2}\frac{(m\dot{x})^2}{m} + V.$$

The Hamiltonian equations of motion are

$$\dot{x} = \frac{\partial H}{\partial p} = \frac{p}{m},$$

and

$$\dot{p} = m\frac{d^2x}{dt^2} = -\frac{\partial H}{\partial q} = -\frac{\partial V}{\partial x} = F.$$

If the Hamiltonian does not depend explicitly on time then it is a constant during the motion; indeed,

$$\begin{aligned}\frac{dH}{dt} &= \sum_{i=1}^{n} \frac{\partial H}{\partial p_i}\frac{dp_i}{dt} + \sum_{i=1}^{n} \frac{\partial H}{\partial q_i}\frac{dq_i}{dt} \\ &= \sum_{i=1}^{n} \frac{\partial H}{\partial p_i}\left(-\frac{\partial H}{\partial q_i}\right) + \frac{\partial H}{\partial q_i}\frac{\partial H}{\partial p_i} \\ &= 0.\end{aligned}$$

The constant value of the Hamiltonian is the energy E of the system. A system of equations which can be put into the form (5.1) is a Hamiltonian system.

5.2. Statistical Mechanics

Consider a Hamiltonian system with n degrees of freedom (q_1, p_1), $\ldots$,(q_n, p_n) where H does not depend explicitly on the time t. From now on we will denote the vector of positions by q and the vector of momenta by p so that $H = H(q, p)$. A microscopic state of the system (a "microstate" for short) is a set of values of the $q_1, \ldots q_n, p_1, \ldots, p_n$.

The system evolves in a $2n$-dimensional space which is denoted by Γ and is often called the phase space. The sequence of points in Γ that the system visits as it evolves from an initial condition is called a trajectory.

If the system has many degrees of freedom then it is impossible to follow its exact evolution in time, since specification of all the initial conditions is impossible and the numerical solution of the very large systems which arise in practice is also out of reach. So we settle for a more modest approach. We assume that the initial data $q(0)$, $p(0)$ are drawn from a probability density W. Then, instead of considering single trajectories we look at the collection, or "ensemble", of trajectories that are initially distributed according to W.

As the trajectories evolve individually the probability density naturally changes; let the density of microstates at time t be $W(t)$, where each microstate is the location of a trajectory at that time. $W(t)$ describes the ensemble at time t; it is the "macrostate" of the ensemble. Thus the microstate is a list of numbers, or a vector in Γ, and the macrostate is a probability density in Γ. The set of all macrostates corresponds to Ω, the sample state of our earlier discussion.

We now derive an equation of motion for $W(t) = W(q, p, t)$. Consider the vector $u = (\dot{q}_1, \ldots, \dot{p}_n)$. First note that its divergence is zero:

$$\begin{aligned}\nabla \cdot u &= \sum_{i=1}^{n} \frac{\partial}{\partial q_i}\left(\frac{dq_i}{dt}\right) + \sum_{i=1}^{n} \frac{\partial}{\partial p_i}\left(\frac{dp_i}{dt}\right)\\ &= \sum_{i=1}^{n} \frac{\partial}{\partial q_i}\left(\frac{\partial H}{\partial p_i}\right) + \sum_{i=1}^{n} \frac{\partial}{\partial p_i}\left(-\frac{\partial H}{\partial q_i}\right)\\ &= 0.\end{aligned}$$

This vector field can be said to be "incompressible", in analogy with fluid dynamics.

Consider a volume V in Γ-space and a density of systems W. The number of microstates in V at a given time t is on the average $\int_V W \, dq \, dp$ (where $dq = dq_1 \ldots dq_n$ and similarly for dp). If microstates neither appear nor disappear, then the only change in the number of systems in V can come from the inflow/outflow of systems across the boundary of V. Therefore, as in fluid mechanics,

$$\frac{d}{dt}\int_V W dq \, dp = -\int_{\partial V} Wu \cdot n \, ds = -\int_V \nabla \cdot (Wu) \, dV,$$

where n is normal to the boundary ∂V of V. If we assume that the density is smooth we can deduce from the above that

$$\frac{\partial W}{\partial t} + \nabla \cdot (Wu) = 0, \tag{5.2}$$

and, using the incompressibility of u,

$$\frac{\partial W}{\partial t} + u \cdot \nabla W = 0. \tag{5.3}$$

This last equation is known as the Liouville equation. One can define a linear differential operator (the Liouville operator)

$$L = \sum_{i=1}^{n} \frac{\partial H}{\partial q_i} \frac{\partial}{\partial q_i} - \sum_{i=1}^{n} \frac{\partial}{\partial p_i} \frac{\partial}{\partial p_i}$$

and then equation (5.3) becomes

$$\frac{\partial W}{\partial t} = -LW. \tag{5.4}$$

This equation is linear even when the original system is not. In as much as it is an equation for the evolution of a pdf, it is analogous to the Fokker-Planck equation; this analogy will be pursued in the next chapter.

Once we have the density $W(t)$, we can define physical observables for the ensemble, which are averages of physical quantities over the ensemble. The energy of each microstate is the value of the Hamiltonian H for that microstate, the energy of the ensemble is

$$E(t) = E[H(t)] = \int_\Gamma H(q,p)W(q,p,t)dV,$$

where dV is an element of volume in the phase space Γ. Similarly, if $\Phi = \Phi(q,p)$ is a property of a microstate, its macroscopic version is

$$\bar{\Phi} = E[\Phi] = \int_\Gamma \Phi(q,p)W(q,p,t)dV.$$

A probability density W is an invariant if it is a stationary solution of equation (5.2), that is, if we draw the initial data from W, solve the equations for each initial datum, and look at the density of solutions at some later time t, it is still the same W. In other words, sampling the density and evolving the systems commute. We now give two examples of invariant densities for a Hamiltonian system.

Suppose that initially W is zero outside a region V and suppose that the system has no way of leaving V. Further suppose that W is constant inside V. Then from equation (5.3) we conclude that W is invariant. We apply this in the following construction. Consider in

Γ-space a surface $H = E$ as well as the surface $H = E + \Delta E$. The volume enclosed between these two surfaces is called an energy shell. Consider the following initial density:

$$W(q,p) = \begin{cases} (\text{volume of shell})^{-1}, & (q,p) \in \text{shell} \\ 0, & \text{otherwise} \end{cases}.$$

Since no systems can leave the energy shell (because the energy is a constant of the motion), this density is invariant. If we let the thickness of the energy shell go to zero, we get a "microcanonical" density. The surface density on the energy surface $H = E$ need not be constant.

Suppose $\phi(H)$ is a function of H such that $\int_\Gamma \phi(H) dqdp = 1$ and $\phi(H) \geq 0$. Thus $W(q,p) = \phi(H)$ is invariant. Note first that $u \cdot \nabla W$ vanishes. Indeed,

$$\begin{aligned} u \cdot \nabla W &= \sum_{i=1}^n \frac{dq_i}{dt}\frac{\partial W}{\partial q_i} + \sum_{i=1}^n \frac{dp_i}{dt}\frac{\partial W}{\partial p_i} \\ &= \frac{\partial \phi}{\partial H}\left(\sum_{i=1}^n \frac{dq_i}{dt}\frac{\partial H}{\partial q_i} + \sum_{i=1}^n \frac{dp_i}{dt}\frac{\partial H}{\partial p_i}\right) \\ &= 0. \end{aligned}$$

Therefore, from equation (5.3), $\partial W/\partial t = 0$. In particular, one can choose as an invariant density $W(q,p) = Z^{-1}\exp(-\beta H(q,p))$, where $\beta > 0$ is a constant and $Z = \int_\Gamma \exp(-\beta H) dq\, dp$. A density of this form is called canonical.

A property of the Liouville operator that will be used later is the following: If $E[\cdot]$ is the expectation with respect to a canonical density, then

$$(Lu, v) = E[(Lu)v] = -E[u(Lv)] = -(u, Lv),$$

i.e., L is skew-symmetric. This can be checked by writing down the definitions and integrating by parts.

5.3. Entropy and Equilibrium

Consider a probability space where Ω consists of a finite number of points $\omega_1, \omega_2, \ldots, \omega_n$ with probabilities $p_1, p_2, \ldots, p_n$ (whose sum must be 1). We now want to define a quantity called "entropy" on that space, to be denoted by S. S will be a function of the p_i: $S = S(p_1, \ldots, p_n)$ and we will consider the case where n may vary. We want S to be a measure of the uncertainty in the probability density and to that end satisfy the following axioms:

(1) For each n, S is a continuous function of all its arguments.

(2) If all the p_i are equal ($p_i = 1/n$ for all i) one can define $S_n = S(1/n, \dots, 1/n)$ and require that S_n be a monotonically increasing function of n (the more points in Ω, the more uncertainty if all points are equally likely).

(3) Let $1 = k_0 \leq k_1 \leq k_2 \leq \cdots \leq k_M = n$ be a partition of $[1, n]$ and let $q_j = p_{k_{j-1}} + \cdots + p_{k_j}$, i.e., $q_1 = p_1 + \cdots + p_{k_1}$, $q_2 = p_{k_1+1} + \cdots + p_{k_2}$, etc. Then

$$S(p_1, \dots, p_n) = S(q_1, \dots, q_M) + \sum_{j=1}^{M} q_j S\left(\frac{p_{k_{j-1}}}{q_j}, \dots, \frac{p_{k_j}}{q_j}\right).$$

In other words, the uncertainty is the sum of the uncertainties inherent in any grouping of points plus the average of the uncertainties within each grouping.

A function S with these properties should be small if all the probability is concentrated at a few points and should become ever larger as there is more doubt as to where an arbitrary point would lie. One can prove that a function S that satisfies these requirements is determined uniquely up to a multiplicative constant and is

$$S = -\sum_i p_i \ln p_i.$$

This is the entropy associated with the probability space we started from. In physics one adds to this definition the multiplicative constant k (Boltzmann's constant). The entropy associated with a pdf f is, similarly, $S = -\int f(x) \ln f(x) dx$. The entropy is a number attached to the pdf which measures, in the way described above, the uncertainty implicit in the pdf.

Now consider a set of microstates (or equivalently, the sample space for an evolving statistical mechanics system), with some reasonable σ-algebra of events. Suppose we have measured some physical, macroscopic, quantities, say $\bar{\Phi}_1, \bar{\Phi}_2, \dots \bar{\Phi}_m$, for some finite m. These are averages with respect to a density W of a set of microscopic (i.e., relating to each state) quantities Φ_i. We now ask the question: What pdf W compatible with these measurements (i.e., such that $\bar{\Phi}_i = \int \Phi_i(q,p) W(q,p) dV$) has maximum entropy? We now show the following: If there exists a vector $\beta = (\beta_1, \dots, \beta_n)$ and a number $Z > 0$ such that

$$W_\beta = Z^{-1} \exp\left(-\sum \beta_i \Phi_i(q,p)\right)$$

is a probability density compatible with the measurements ("admissible" for short) then W_β is the admissible density that has the largest entropy among all admissible densities.

The proof is as follows. It is an exercise in calculus to show that $\psi(x) = x \ln x - x + 1 \geq 0$ for $x \geq 0$, with equality only for $x = 0$. Put $x = W/W_\beta$ in this inequality, where W is an arbitrary admissible density. Then

$$-W \ln W + W \ln W_\beta \leq W_\beta - W.$$

Integrate this inequality over Γ, and use the fact that both W and W_β are densities; this gives

$$-\int_\Gamma W \ln W dV \leq -\int_\Gamma W \ln W_\beta \, dV.$$

However, from the definition of W_β we find that

$$-\int_\Gamma W \ln W_\beta \, dV = -\int_\Gamma W_\beta \ln W_\beta \, dV = \ln Z + \sum \beta_i \bar{\Phi}_i$$

so that all the entropies of the W's are less than the entropy of W_β:

$$S(W) \leq S(W_\beta),$$

where $S(W)$ is the entropy associated with a density W. Furthermore, the inequality is strict unless $W = W_\beta$.

As an example, suppose one has a single measurement, that of E, the energy of the ensemble, $E = E[H]$; then $W_\beta = Z^{-1}e^{-\beta H}$, where the β in the exponent is a scalar, and $Z = \int_\Gamma e^{-\beta H} dV$. The parameter β is determined from the equation

$$E = E[H] = \int_\Gamma Z^{-1} H e^{-\beta H} dV = \frac{\partial Z}{\partial \beta}.$$

With this density, the entropy is $S = \beta E + \ln Z$. There is a calculation, which we omit, producing the microcanical density in the absence of any measurements.

It is a physical principle that the entropy of a physical system always increases, so it is reasonable to assume that any density for a physical system will evolve in time into one that maximizes the entropy. We already know that a canonical density is time invariant, so the canonical density is a good candidate for an asymptotic, invariant density, what is called in physics a "thermal equilibrium." This is particularly satisfying from the point of view of statistics as well: one can show that estimates based on partial measurements are unbiased if one assumes that the density that gives rise to them maximizes the entropy.

The temperature T of a system is defined by the equation

$$T^{-1} = \partial S / \partial E,$$

one can check that if the density is the canonical density above then $T = 1/\beta$ (in physics there is an additional factor of k from the physicists' definition of entropy). Then the canonical density can be written as $W = Z^{-1}\exp(-H/T)$. For a system of N non-interacting particles, T/N can be seen to be the variance of the velocity of each particle divided by its mass. The canonical density has T as a fixed parameter, and is the right density to use when the system under study allows no exchange of mass through its walls and has walls kept at a fixed temperature T. For the sake of simplicity, in these notes we shall always place ourselves in this case.

One can now proceed to derive all of thermodynamics from our definitions but we forbear to do so. We merely pause to note that the normalization constant Z varies when T varies, and is known in physics as the "partition function."

Suppose $F = F(q,p)$ is a function on Γ and suppose you want to calculate the average of F along a trajectory of a system in thermal equilibrium, e.g., for a system that has an invariant measure (such as the measure defined by the canonical density). If the "ergodic property" holds, this average equals the average of F with respect to the invariant measure. If one can prove the ergodic property (as one very occasionally can), or assume it holds (as one often does) then the calculation of averages is greatly simplified. An example of an ergodic system is the system where Γ is the interval $[0,1)$, and the equation of motion is $x_n = (x_{n-1}+\gamma) \bmod 1$, with x_0 given. One can readily check that if γ is irrational, then the standard Lebesgue measure on $[0,1)$ is invariant, and that the average of any continuous function F defined on $[0,1)$ with respect to Lebegues measure equals its average over any trajectory.

5.4. The Ising model

We now introduce the Ising model in two space dimensions, which is widely used as a model problem in statistical mechanics. Consider an $N \times N$ regular lattice in the plane with lattice spacing 1, and at each node (i,j) set a variable $s_{i,j}$ (a "spin") that can take only one of two values: $s_{i,j} = 1$ ("spin up") or $s_{i,j} = -1$ ("spin down"). Make the problem periodic, so that $s_{i+N,j} = s_{i,j}$ and $s_{i,j+N} = s_{i,j}$. Associate with this problem the Hamiltonian

$$H = -\sum s_{i,j}(s_{i+1,j} + s_{i-1,j} + s_{i,j+1} + s_{i,j-1}),$$

i.e., minus the sum of the products of each spin with its four nearest neighbors. This "Hamiltonian" does not include any momenta, and the variables take integer values only, so there is no time evolution associated with it and "thermal equilibrium" here is meaningful only in the sense that the probability density we use maximizes the entropy.

The possible microstates of the system are the 2^{N^2} ways of arranging the up and down spins. We assign to each microstate the probability $Z^{-1}\exp(-H/T)$, where as above T is the temperature. A function of the microstate that is of interest is the "magnetization"

$$\mu = \frac{1}{N^2}\sum_{i,j} s_{i,j}.$$

Clearly if all the spins are aligned $\mu = +1$ or $\mu = -1$. With the definitions above, $E[\mu] = 0$ because a state with a given set of values for the spins and a state with exactly the opposite values have equal probabilities.

The covariance function is

$$\mathrm{Cov}(i', j') = E[s_{i,j}s_{i+i',j+j'}].$$

The correlation length is a number ξ such that for $\|(i', j')\| = \sqrt{i'^2 + j'^2} > \xi$ the covariance is not significant (and we do not explain further how big "significant" is).

One can show, and check numerically as explained below, that the Ising model has the following properties:

(1) For T very large or very small ξ is small, of the order of 1. There is an intermediate value T_c of T for which ξ is very large.

(2) The behavior of the magnetization μ is very different when $T < T_c$ and when $T > T_c$. In the former case the likely values of μ hover around two non-zero values $\pm\mu_*$; if one adds dynamics to this problem (as we shall do with Monte-Carlo sampling in the next section) one sees that the system is very unlikely to move from $+\mu_*$ to $-\mu_*$ or vice-versa. For very large values of N the phase space Γ separates into two mutually inaccessible regions which correspond to μ positive and μ negative. The averages of μ over each region then have one sign. On the other hand, when $T > T_c$ this separation does not occur. The value $T = T_c$ is a "critical value" of T and the parameter m is an "order parameter" which can be used to detect the partial order in which spins are aligned in each of the two mutually inaccessible regions of Γ. As T passes from above this value T_c to below the critical value T_c one has a "phase transition" in which the system goes from a disordered "phase" to a partially ordered phase. If one averages μ for $T < T_c$ only over the appropriate part of the phase space, one finds that

when $|T - T_c|$ is small, m is proportional to $|Tc - T|^\alpha$, where $\alpha = 1/6$ is an instance of a "critical exponent."

5.5. Markov Chain Monte Carlo

Let $\phi(q,p)$ be a scalar function of the q's and p's, i.e., $\phi{:}\Gamma \to \mathbb{R}$. We want to compute the expectation value of ϕ with respect to the canonical density:

$$E[\phi] = \int_\Gamma \phi(q,p) \frac{e^{-H(q,p)/T}}{Z} dq\, dp.$$

The estimation of such integrals is difficult because: i) usually the number of variables is large, ii) the partition function Z is unknown, and iii) $H(q,p)$ is usually very small except on a very small part of Γ, so that without some form of importance sampling the computation takes forever.

An excellent method for calculating such integrals is "Markov chain Monte Carlo" or "Metropolis sampling" or "rejection sampling", which will now be explained. To simplify the analysis, consider a system with a finite number of microstates $S_1, S_2, \ldots, S_n$. To each microstate we assign a value $H_i = H(S_i)$ of the Hamiltonian and a probability

$$P_i = P(S_i) = \frac{e^{-H_i/T}}{Z}, \tag{5.5}$$

where

$$Z = \sum_{i=1}^{n} e^{-H_i/T}.$$

Suppose $\phi = \phi(S)$ is a function on the space $\Gamma = \{S_1, \ldots, S_n\}$. We have

$$E[\phi] = \sum_{i=1}^{n} \phi(S_i) P_i = \sum_{i=1}^{n} \phi(S_i) \frac{e^{-H_i/T}}{Z}.$$

Definition. Consider a space Γ containing states $S_1, S_2, \ldots, S_n$. A time series on Γ (or a chain on Γ) is a time series X (discrete time stochastic process, see Chapter 4) such that for each instant t, $X_t = S_j, 1 \le j \le n$.

Example. Suppose the space Γ consists of the states S_1, S_2, S_3, S_4. A sequence of states $X_t = (S_1, S_1, S_3, S_2, \ldots)$ is a time series.

Definition. The probability

$$P(X_t = S_j | X_{t-1} = S_{j_1}, X_{t-2} = S_{j_2}, \ldots)$$

is called the transition probability of the chain and the chain is a Markov chain if

$$P(X_t = S_j | X_{t-1} = S_i, X_{t-2} = S_{i_2}, \ldots) = P(X_t = S_j | X_{t-1} = S_i).$$

For a Markov chain we write

$$P(X_t = S_j | X_{t-1} = S_i) = p_{ij} = P(S_i \to S_j),$$

where $\sum_j p_{ij} = 1$ and $p_{ij} \geq 0$. The matrix $\mathbf{P}$ with elements p_{ij} is called the Markov matrix.

Suppose that we know $P(S_i \to S_j) = p_{ij}$. We have

$$\begin{aligned} P(X_t = S_j | X_{t-2} = S_i) &= \sum_k P(S_i \to S_k) P(S_k \to S_j) \\ &= \sum_k p_{ik} p_{kj} \end{aligned}$$

which are the entries of the matrix $\mathbf{P}^2$. If $\mathbf{P}^{(2)}$ is the matrix whose entries are the probabilities that we go from S_i to S_j in two steps, then $\mathbf{P}^{(2)} = \mathbf{P}^2$.

Definition. A collection of states is irreducible with respect to a chain X if given any two states S_i, S_j in the collection (where we may have $i = j$) there is a non-zero probability of going from S_i to S_j in n steps for some n.

Definition. A Markov chain is ergodic in Γ if all of Γ is irreducible with respect to the chain.

Example. Consider the four state system above and let X be the finite chain $\{S_1, S_2, S_1\}$. For simplicity we assume that $\mathbf{P}$ has all non-zero entries. In this case X is not irreducible with respect to $\{S_1, S_2\}$ because the transition $S_2 \to S_2$ does not occur. However, $\{S_1, S_2\}$ is irreducible with respect to $X = \{S_1, S_2, S_2, S_1\}$. Neither of these chains is ergodic in Γ. An example of an ergodic chain is $\{S_1, S_2, S_3, S_4, S_4, S_3, S_2, S_1\}$.

The following theorem holds.

Theorem 5.1. *If a Markov chain is ergodic in Γ then there exist numbers π_i such that $\pi_i \geq 0$, $\sum_i \pi_i = 1$, and $\pi_j = \sum_i \pi_i p_{ij}$.*

The probabilities $\{\pi_i\}$ are the analog of an invariant density when the system is Hamiltonian.

For the sake of simplicity we work out the next steps in the special case of a one dimensional Ising model, in which nothing interesting happens and there is no phase transition. The n spins now live on a one-dimensional lattice. The Hamiltonian H associated with a microstate is

$$H = -\sum_{i=1}^{n} s_i s_{i+1},$$

where, as before, the domain is periodic so that $s_{i+n} = s_i$. The probability density on the space of all possible configurations is given by (5.5). Let $\phi(S)$ be any function of the configuration S. Then $\phi(S)$ is a random variable. Fix a specific configuration S_j. Of particular interest is the random variable $\phi(S) = \delta_j(S)$ defined to be 1 if $S = S_j$ and zero otherwise—the characteristic function of S_j . Let X_t be a Markov stochastic process of length N with the values in the space of configurations of an Ising chain. Define

$$\Delta_j(N) = \frac{1}{N}\sum_{t=1}^{N} \delta_j(X_t).$$

Let π_j be the limit of $\Delta_j(N)$ as $N \to \infty$. Then π_j is the frequency with which the chain visits configuration S_j. If the Markov chain X_t is ergodic there exist unique numbers π_j as in theorem 5.1. Additionally in this case π_j is the limit of $\Delta_j(N)$ as $N \to \infty$.

Suppose we know how to find p_{ij} so that the $\pi_j = Z^{-1}e^{-H_j/T}$ are the predetermined probabilities; then if $\phi(S)$ is any function on the configuration space of the one dimensional Ising lattice and if X_t is ergodic we have

$$\frac{1}{N}\sum_{t=1}^{N} \phi(X_t) \to E[\phi(S)].$$

We have to be very careful in designing the Markov chain X_t or we will mostly do useless work.

Example. Consider a one dimensional Ising model with 4 sites. There are $2^4 = 16$ possible configuration of the chain; for instance, one possible configuration (or microstate) is $S = (+1, -1, -1, +1)$. The possible values of the Hamiltonian are $-4, 0, 4$. There are two states with $H = -4$ (these are the states for which all s_i's are of the same sign), 12 states with $H = 0$, and two states with $H = 4$ (the states with alternating signs). Suppose the temperature is $T = 1$, then, using (5.5), the two states with all s_i's of the same sign each have probability of about 0.45. Together they have probability 0.9 of appearing. The next most likely state has a probability of only 0.008. The situation becomes even more dramatic as the number of sites in the Ising lattice increases. In general there will be a very small number of states with significant probabilities and a very large number of states with probabilities near zero. Thus if we want to compute the average of some random variable $\phi(S)$ it would not make sense to sample each site with equal frequency. We must construct a chain which visits the sites with probability equal to

$$\pi_i = \frac{1}{Z} e^{-H_i/T}.$$

The change in sampling to reach this goal is what we called importance sampling in Chapter 2.

We construct Markov processes that accomplish importance sampling in two steps. First we do something stupid, then we cleverly improve it.

Step 1. We construct an arbitrary ergodic symmetric Markov chain (a Markov chain is symmetric if $p_{ij} = p_{ji}$). For example, in the Ising case, we start our chain with an arbitrary configuration. At each time step we pick a number i between 1 and n with equal probability and change the value s_i associated to site i to the opposite value: $s_i \to -s_i$.

Step 2. Suppose the Markov process defined above has transition probabilities p_{ij}. We construct a modified Markov chain by defining a new transition probabilities p^*_{ij}.

Case $i \neq j$. In this case

$$p^*_{ij} = \begin{cases} p_{ij}\dfrac{\pi_j}{\pi_i}, & \dfrac{\pi_j}{\pi_i} < 1 \\ p_{ij}, & \dfrac{\pi_j}{\pi_i} \geq 1 \end{cases}.$$

Case $i = j$. In this case

$$p^*_{ii} = p_{ii} + \sum p_{ij}\left(1 - \frac{\pi_j}{\pi_i}\right)$$

where the sum is over all j such that $\pi_j/\pi_i < 1$.

We claim that the modified process visits configuration S_j with probability π_j. This is a consequence of the fact that $\sum_j p^*_{ij}\pi_i = \pi_i$.

How to apply this result: Let P be the transition matrix of some ergodic Markov process on the states $\{S_j\}$. Suppose that we are currently in the state S_i. We use $\mathbf{P}$ to pick the next state S_j, the transition probability of this is p_{ij}. Having picked S_j in this way we calculate the ratio π_j/π_i. If $\pi_j/\pi_i \geq 1$ we accept S_j as the new state. On the other hand if $\pi_j/\pi_i < 1$, then with probability π_j/π_i we accept S_j as the new state , and with probability $1 - \pi_j/\pi_i$ we take the old state S_i to be the new state. This procedure gives the transition probabilities p^*_{ij} defined above.

A very important observation is that

$$\frac{\pi_j}{\pi_i} = \exp\left(-\frac{H(S_j)}{T} + \frac{H(S_i)}{T}\right) = \exp\left(-\frac{\Delta H}{T}\right),$$

where ΔH is the difference in energy between the states S_i and S_j. Note that Z is never needed.

This construction is easy to program and quite efficient in general. The exception is in more than one space dimension for T near the critical value T_c. The problem is as follows: We have seen that the error in Monte-Carlo methods depends on the number of samples used, and was estimated on the assumption that these samples were independent. If the samples are not independent more samples are needed. Near T_c the spatial correlation length is very large, and so is the temporal correlation time of the Monte-Carlo samples—more and more Metropolis moves are needed to obtain a spin configuration independent of the previous one, and the cost of the calculation diverges (this is known as "critical slowing-down"). A cure will be described in the next section.

5.6. Renormalization

Consider again a Hamiltonian system like those at the beginning of this chapter: $2n$ equations for n pairs q_i, p_i, $i = 1, \ldots, n$, satisfying the ordinary differential equations $\dot{q} = \partial H/\partial p$, $\dot{p} = -\partial H/\partial q$, with $H = H(q,p) > 0$. If the initial data are sampled from the density $Z^{-1}\exp(-H/T)$ these equations can be used to calculate averages of smooth functions ϕ with respect to this initial density (which is invariant). Alternatively, such averages can be calculated by Markov chain Monte-Carlo.

Now suppose the functions we wish to average depend on a subset of the variables q, p, say only on the variables $q_1, q_2, \ldots, q_m, p_1, \ldots, p_m$, with $m < n$; denote this set of components by $\hat{q}$, $\hat{p}$. Is it possible to sample the values of $\hat{q}$, $\hat{p}$ without sampling the others, or solve a system of equations for these "resolved" variables which does not involve the others?

The equations of motion for the components of $\hat{q}, \hat{p}$ involve all the components of q, p (or else the question has already been answered in the affirmative). Approximate the right-hand-side of these equations by their best approximation by a function of the $\hat{q}, \hat{p}$ we wish to calculate, following the time-honored expedient of trying to approximate a solution by first approximating the equations; this yields the system of equations

$$\frac{d}{dt}\hat{q}_i = E\left[\left.\frac{\partial H}{\partial p_i}\right| \hat{q}, \hat{p}\right], \quad \frac{d}{dt}\hat{p}_i = E\left[\left.-\frac{\partial H}{\partial q_i}\right| \hat{q}, \hat{p}\right],$$

for $i \leq m$.

We make the following claims (these are the "Hald theorems"):

(1) This new, reduced system for the $2m$ resolved variables is also Hamiltonian. Let the set of components of q not in $\hat{q}$ be denoted by $\tilde{q}$, and similarly for $\tilde{p}$. Let $i \leq m$, so that q_i is in $\hat{q}$. Then

$$E\left[\left.\frac{\partial H}{\partial p_i}\right| \hat{q}, \hat{p}\right] = \int \frac{\partial H}{\partial p_i} e^{-H/T} d\tilde{q}\, d\tilde{p}$$

by definition of the conditional expectation (see Chapter 2); $d\tilde{q}$ denotes integration over all the components of $\tilde{q}$, and the same for $d\tilde{p}$. The last expression equals $\partial\hat{H}/\partial p_i$ where

$$\hat{H} = -T \int e^{-H/T} d\tilde{q}\, d\tilde{p}.$$

A similar identity holds for the p variables. $\hat{H}$ is a new Hamiltonian, the "renormalized" Hamiltonian with respect to the partition of q, p into $\hat{q}, \hat{p}$ and $\tilde{q}, \tilde{p}$ (the name comes from applications in quantum theory).

(2) The normalization constant $\hat{Z}$ (the partition function) for the new Hamiltonian equals the normalization constant for the original Hamiltonian, $\hat{Z} = Z$, as can be checked from the definitions.

(3) The density $\hat{W} = Z^{-1}\exp(-\hat{H}/T)$ is invariant for the reduced system (for the same reasons that $W = Z^{-1}\exp(-H/T)$ is invariant for the old system) and $\hat{W}$ the joint density of the resolved variables, equals their marginal density in the old system (i.e., the density W integrated over all the variables not in the resolved set). This is easy to see:

$$\begin{aligned} Z^{-1}e^{-\hat{H}/T} &= Z^{-1}\exp\left(-\int e^{-H/T} d\tilde{q}\, d\tilde{p}\right) \\ &= Z^{-1}\int e^{-H/T} d\tilde{q}\, d\tilde{p}. \end{aligned}$$

If one wants to average only over the resolved variables one can solve the reduced system in time, or sample the canonical density associated with the renormalized Hamiltonian by Markov chain Monte-Carlo.

We now apply these idea to the Ising model; for simplicity we write things for the one-dimensional Ising model even though this model is of no physical interest in one dimension; to make the notations consistent we write s in place of the q of the preceding paragraphs.

First, note that the spins take discrete values and differentiation of H with respect to the s_i is not obviously well defined. One can extend the range of the spins in many ways, for example make them take values in $\mathbb{R}$, and then make the Hamiltonian $Z^{-1}\exp(-(H+K)/T)$, where $K=\epsilon^{-1}\Pi_i(s_i-1)(s_i+1)$. For small ϵ this forces the spins to take values near $+1$ or -1. One can proceed with the analysis below with ϵ small but finite (these are "soft" spins), and at the end of the analysis make ϵ tend to zero so the spins become Ising spins.

Second, there are no momenta in the Ising Hamiltonian and no time dependence. However, one readily checks that the Hald theorems hold if one defines the renormalized Hamiltonian $\hat{H}$ by

$$\frac{\partial \hat{H}}{\partial s_i} = E\left[\left.\frac{\partial H}{\partial s_i}\right| \hat{s}\right]$$

for all $i \leq m$. Furthermore, the Ising model is translation invariant. All the constructions below will be translation invariant as well, so this last equation will be satisfied for all i if it satisfied for one of them, say for $i=1$.

The Hamiltonian for the Ising system has terms of the form $s_i s_{i'}$, where the points i and i' are neighbors; we say that the Hamiltonian "couples" nearest neighbors; nearest neighbors appear together in it (but of course the solution has interactions among non-nearest neighbors). A Hamiltonian of the form $\sum_i s_i s_{i+j}$, for j fixed, is said to have a coupling between spins j apart.

To evaluate $\hat{H}$, start by considering a Hamiltonian more general that the Ising Hamiltonian, of the form:

$$H = a_1 H_1 + a_2 H_2 + a_3 H_3 + \dots, \tag{5.6}$$

where the a_i are constants and each "sub-Hamiltonian" H_j has only couplings between spins j apart; one could have for example $H_j = \sum_i s_i s_{i+j}$ plus powers of this last sum. Assume at first that one has only the simplest quadratic terms just written out—these will not be sufficient in real life but they will suffice for the explanation here. There is no need to worry about convergence, because to start with the number of terms will be finite; indeed, if one picks H_1 as H, the spin Hamiltonian, then equation (5.6) is an identity for our H with $a_1=1$ and all the other $a_i=0$. Now differentiate (5.6) with respect to s_1, assumed to be part of $\hat{s}$, i.e., a variable that will be kept:

$$\frac{\partial H}{\partial s_1} = \sum a_j \psi_j(s),$$

where $\psi_j(s) = \partial H_j / \partial s_1$. Applying the projection operator $\mathbb{P}$ defined by $\mathbb{P}g = E[g|\hat{s}]$ to both sides, we find

$$\frac{\partial \hat{H}}{\partial s_1} = \sum_j \mathbb{P}\psi_j. \tag{5.7}$$

With the conventions above, where $\hat{s}$ included the s_i with i odd and $\tilde{s}$ the s_i with i even, ψ_j for j even depends only on variables in $\hat{s}$, so that the projection leaves them invariant. For j odd, one has to calculate the inner products of ψ_{odd} with the ψ_{even} and proceed to find the projection as in Chapter 1; all the series converge as shown there. Collecting terms, one finds

$$\frac{\partial \hat{H}}{\partial s_1} = \sum_{j \text{ even}} \hat{a}_j \psi_j,$$

where the $\hat{a}_j$ are the coefficients found by collecting terms. It follows that

$$\hat{H} = \sum_{j \text{ even}} \hat{a}_j H_j.$$

The price for reducing the number of variables is an increase in the complexity of the Hamiltonian. One more step is needed: we now have $\hat{H}$ as a function of the $\hat{s}_i$ for odd values of i only. Reposition these variables so that $s_i, i = 2i^* + 1$ moves to s_{i^*}. We now have spins at the same locations as previously. The transformation which consists in first doing $(q,p) \to (\hat{q},\hat{p})$, $H \to \hat{H}$, followed by repositioning, is an instance of a "real-space", or Kadanoff, renormalization group (RNG) transformation. RNG transformations are used to simplify calculations all over physics. If the reduction after one RNG transformation is not sufficient, one can proceed recursively: define $H^{(0)} = H$, $H^{(1)} = \hat{H}$, $H^{(2)} = \hat{H}^{(1)}, \ldots$ etc.

What has been gained? In one dimension, computationally, not much. The number of variables has been decreased but the Hamiltonian has become more complex, and this looks like a wash. In two (or more) dimensions, a lot can be gained. If T is far from the critical temperature, the correlations length ξ is small and a good approximation to the statistics of a large array of spins should be computable on a small array of spins. However, near T_c one cannot cut off the array size to something small without distorting the correlations; one can in addition expect critical slowing down so Monte-Carlo calculations can be very expensive. However, each time one renormalizes, the correlation length is reduced by a factor which depends on the size of the boxes (by 1/2 in the example above), and the temporal correlations also decrease, so the renormalization takes one away from T_c and makes computation easier.

For $T = T_c$, ξ is actually infinite; halving infinity still yields infinity. The RNG transformations have a fixed point at T_c, and their analysis near T_c also yields the critical exponents; this topic exceeds the scope of the present notes. Note an analogy between these remarks and the central limit theorem: Suppose you have a collection of random variables and the pdf of their sum is exactly Gaussian; the pdf of a subsum will still be Gaussian, but if the pdf of the sum is only approximately Gaussian, the pdfs of subsums will be ever more different from Gaussian when one makes the subsets smaller, as one can see by reversing the process of adding variables and averaging that leads to the central limit theorem. In this sense, RNG theory is a generalization of the central limit theorem to a case where the variables are dependent.

5.7. References

1. G.I. Barenblatt, *Scaling, Self-Similarity, and Intermediate Asymptotics*, Cambridge U. Press, Cambridge, ((1996)
2. A.J. Chorin, *Conditional expectations and renormalization*, Multiscale Modeling and Simulation, 1, (2003), pp. 105-118.
3. H. Dym and H. McKean, loc. cit.
4. L.C. Evans, *Entropy and Partial Differential Equations*, lecture notes, UC Berkeley Math. Dept., 1996.
5. N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group*, Perseus, Reading, Mass, (1992)
6. J. Hammersley and D. Handscomb, *Monte-Carlo Methods*, Methuen, London, (1964).
7. E.T. Jaynes, *Papers on Probability, Statistics and Statistical Physics*, Reidel, (1983).
8. G. Jona-Lasinio, *The renormalization group- a probabilistic view*, Nuovo Cimento, 26 (1975), pp. 99-118.
9. L. Kadanoff, *Statistical Physics: Statics, Dynamics, and Renormalization*, World Scientific, Singapore, (1999).
10. D. Kandel, E. Domany and A. Brandt, *Simulation without critical slowing down- Ising and 3-state Potts model*, Phys. Rev. B 40 (1989), pp. 330-344.

11. C. Thompson, *Mathematical Statistical Mechanics*, Princeton U. Press, Princeton NJ, (1972).